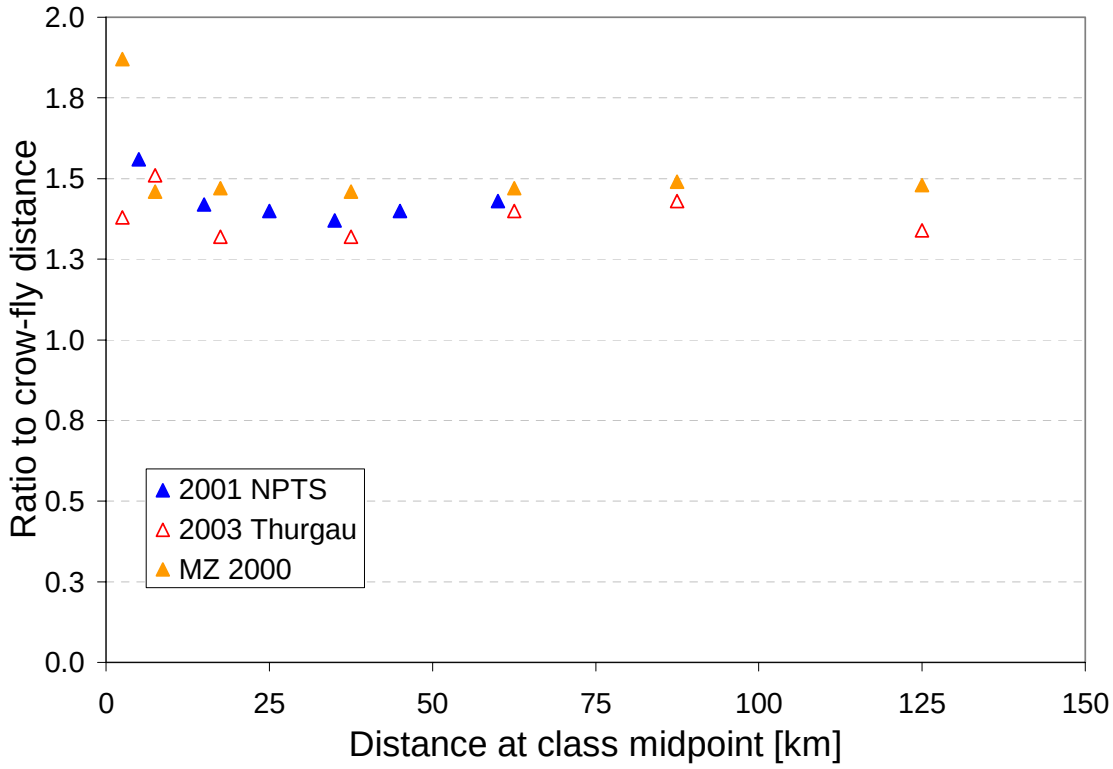


Precision of geocoded locations and network distance estimates

VS Chalasani

Ø Engebretsen

JM Denstadli

KW Axhausen

Arbeitsbericht Verkehrs- und Raumplanung 256 **September 2004**

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Arbeitsberichte Verkehrs- und Raumplanung

Genauigkeit von geokodierten Standorten und Netzwerkdistanzen

VS Chalasani
IVT
ETH Zürich
CH – 8093 Zürich

Telephon: +41-1-633 xxxx
Telefax: +41-1-633 1057
chalasani@ivt.baug.ethz.ch

Ø Engebretsen
JM Denstadli
TØI
P.O. Box 6110
Etterstad
N-0480 Oslo

Telephon: +47-22-57 38 00
Telefax: +47-22-57 02 90
jmd@toi.no

KW Axhausen
IVT
ETH Zürich
CH – 8093 Zürich

Telephon: +41-1-633 3943
Telefax: +41-1-633 1057
axhausen@ivt.baug.ethz.ch

September 2004

Kurzfassung

Dieser Aufsatz präsentiert die Ergebnisse von Auswertungen der norwegischen nationalen Verkehrsbefragung 2001, des Schweizer Mikrozensus Verkehr 2000 und der 6-Wochen Tagebuchbefragung im Thurgau 2003 zu den Themen: Welche Genauigkeiten sind heute bei der Gekoiderung von Wegen erreichbar ? Gibt es systematische Differenzen zwischen den verschiedenen netzwerk-gestützten Entfernungsschätzungen ? Wie brauchbar sind die berichteten Entfernungen der Befragten ?

Die Ergebnisse zeigen, dass heute die überwiegende Mehrheit aller Wege sehr präzise lokalisiert werden können, wenn die Befragung sich dieses Ziel setzt. Die verschiedenen Entfernungsschätzungen haben vorhersagebare und stabile Zusammenhänge miteinander. Das gilt auch für die berichteten Entfernungen, die zu mindest im Mittel für statistische Zwecke verwendbar sind.

Schlagworte

Geokodierung, Tagebücher, Wege, Distanzen, Netzmodelle, Genauigkeit

Zitierungsvorschlag

Chalasani, V.S., J.M. Denstali, Ø. Engebretsen und K.W. Axhausen (2004) Precision of geocoded locations and network distance estimates, *Arbeitsbericht Verkehrs- und Raumplanung*, **256**, IVT. ETH Zürich, Zürich.

Working Paper

Precision of geocoded locations and network distance estimates

VS Chalasani
IVT
ETH Zürich
CH – 8093 Zürich

Telephone: +41-1-633 3340
Telefax: +41-1-633 1057
chalasani@ivt.baug.ethz.ch

Ø Engebretsen
JM Denstadli
TØI
P.O. Box 6110
Etterstad
N-0480 Oslo

Telephone: +47-22-57 38 00
Telefax: +47-22-57 02 90
jmd@toi.no

KW Axhausen
IVT
ETH Zürich
CH – 8093 Zürich

Telephone: +41-1-633 3943
Telefax: +41-1-633 1057
axhausen@ivt.baug.ethz.ch

September 2004

Abstract

This paper addresses three questions: how accuraate is the geocoding of travel diaries; what are the relationships between different network-based distance estimates, and how exact are estimates provided by self-reported distances.

Three large-scale surveys in Norway and Switzerland demonstrate that very high precision is possible when survey protocol emphasises capture of addresses. The necessary databases and networks are available today. Crow-fly, shortest-distance path, shortest-time path, and mean UE path distances are systematically related to each other, the pattern of their relationships matching theoretical expectations and the resolution of the networks used. In the examples studied, medians of self-reported distances by distance band provide reasonable estimates of crow-fly and shortest-distance path distances.

Keywords

Geocoding, travel diary, precision, network distances, detour factors

Preferred citation style

Chalasani, V.S., J.M. Denstali, Ø. Engebretsen and K.W. Axhausen (2004) Precision of geocoded locations and network distance estimates, *Arbeitsbericht Verkehrs- und Raumplanung*, **256,** IVT. ETH Zürich, Zürich.

1 How much precision is possible ?

Measuring distances travelled is a central task of transport statistics, as this number is not only a key descriptor of travel behaviour, but also essential for the calculation of derived statistics, such as exposure to risks (accidents, exhausts), volume of externalities (emissions, congestion), speeds, incidence of taxation, etc.. It is also central, directly or indirectly, to all choice models estimated from travel behaviour data. Thus, it is not surprising that recent technological innovations, such as geographic information systems and the vast expansion of spatially referenced data bases and networks have been adopted quickly by transport statisticians and modellers. This adoption process is still ongoing, and professional standards for appropriate use still must be formulated. This paper contributes to the ongoing discussion; first, by highlighting various questions about the availability of these new resources and second, by reporting results from our work with them in Norway and Switzerland.

The gold standard of distance measurement is an uninterrupted trace of GPS points matched to a complete and geometrically correct network model. The currently available GPS data sets are neither uninterrupted, nor matched to complete and geometrically correct network models (See for a recent example Hackney, Marchal and Axhausen, 2004 or Marchal, Hackney and Axhausen, 2004), but they are much closer to this standard than the alternatives discussed below. Some studies actually come quite close; see, for example, Wolf, Oliveira and Thompson, 2003. Lacking data of this quality, the researcher has various second-best alternatives to locate (geocode) origins and destinations of stages or trips observed (Axhausen, 2003), and to estimate distances between them. Data sources assumed available for further discussion are; travel diary surveys (see Richardson, Ampt and Meyburg, 1995; Axhausen, Madre, Polak and Toint, 2003 and Resource Systems Group, 1999), address data bases, and network models suitable for shortest path calculations.

Quality of geocoding will depend on details reported by travellers, as well as detail of the address database to which these reports are matched. Travellers' difficulties with reporting addresses are well known: full street addresses are not known for shops and other locations; correct post codes are forgotten, even when the street address is known; or no unique names exist for common meeting points in parks or other public spaces. Address data bases have similar problems: no entries for points in public spaces, arbitrary allocation of reference points for large complexes, such as stations, airports, or shopping centres, and some missing street addresses.

Using zones for modelling convenience or privacy protection increases both complexity and the possibility of error. . The definition of a reference point for a zone is an additional problem in its own right. Should one use the geographical mean of the zone, of the built-up area, or the centre of gravity of the population, or the city hall, or the post office, for post code – defined zones ?

Currently available detailed network models for vehicle navigation are almost perfect from a topological perspective, as they include (nearly) all street addresses and all nodes. Minor delays in the updating of such databases also cause only minor errors. The larger issue is the coding of link types and associated mean speeds for link types. The same problems (with larger impacts on accuracy) occur with planning networks, i.e. networks used in planning applications for assignment or other transport flow algorithms (Ortuzar and Willumsen, 2001 or Sheffi, 1985). These contain far fewer links and nodes, causing differences between shortest paths calculated using them in comparison with using navigation networks. An added complication is their use of zones to represent space with all the related definition problems discussed above. In addition, network models employ special types of links to connect zones with networks. One such connector is required to produce a complete description of the area, but many users employ two or more, which again will impact shortest path calculations.

Road geometry in network models only approximates the true geometry of real road alignments. As long as the true length of links is known, locating a street address along a link will add only minor errors.

Network models can be used to calculate path distances between origins and destinations for different criteria, which might or might not have the same values, for example:

- Shortest-distance path
- Time-shortest path
- Paths included in the set of paths travelled at user-equilibrium
- Paths included in the set of paths travelled at stochastic user equilibrium
- Paths included in the set of paths travelled at system optimum

For the last three criteria, one would need to define summaries of returned path distances, e.g., mean, median, or minimum. The complexities involved in estimating origin-destination matrices required for these calculations will be ignored here, but see Ortuzar and Willumsen, 2001 for details.

Calculation of shortest distance path distance is unambiguous. This is not the case for shortest-time path distance, which requires the modeller to make assumptions about travelling

speeds on the various links.. One obvious assumption is the free-flow speed, normally the posted speed limit, available in all assignment networks. Most networks set up for navigation purposes assume a mean speed for each link type. These are substantially lower than free flow speeds. Other a-priori choices are possible.

One can also calculate the straight line (crow-fly) distance between two points, either as Euclidian distance or as Great Circle distance (Hubert, 2003), that takes the Earth's spherical shape into account.

When one considers the number of possible combinations and choices in network distance calculaion, traveller-reported distances are at least unambiguous, even if generally inaccurate. Travellers' inability to estimate distances is well known (See Bovy and Stern, 1990; Rietveld, Zwart, van Wee and van der Hoorn, 1999 or Raghubir and Krishna, 1996). In many cases, though, this is the only information available. Thus, patterns of deviations between reported and modelled distances are of interest.

Although not yet undertaken, a study of the interactions between all these elements would be interesting. This paper will focus on many of these relevant issues that provide some missing background allowing other results to be assessed:

- What degree of accuracy is possible in the geocoding of addresses obtained from travel diaries? The results of three studies, the Swiss national travel diary survey (Mikrozensus 2000), the 2003 Thurgau six-week diary (Thurgau 2003), and the 2001 Norwegian national passenger travel survey (NPTS 2001) will be compared.
- How large are the differences between various distance estimates? Using a current national assignment model for Switzerland (Vritc, Fröhlich and Axhausen, 2003 or Vritc and Axhausen, 2004), distance-shortest path distance, time-shortest-path distances, and mean user equilibrium path distances will be calculated and compared.
- What are the differences between reported distances and calculated distances? The three datasets will be used to answer this question.

The structure of the paper will follow the sequence of these questions, but the next section will introduce the surveys. Conclusions and a discussion of future research questions are contained in the final section.

2 Datasets

2.1 2001 Norwegian National Passenger Travel Survey (2001 NPTS)

The 2001 NPTS is the latest in a series of Norwegian travel surveys, which are undertaken on a four year cycle (Denstadli, Hjorthol, Rideng and Lian, 2003). The respondents, all of whom are at least 13 years old, report both their trips for one day, and all trips over 100km made during the last month in a computer-aided telephone (CATI) interview. They had been asked to fill in a 'memory jogger' before the interviews. Respondents are drawn from the national person register, which allows a pre-geocoding of home and work place addresses.

The published data set gives addresses at the level of the approximately 14'000 statistical wards, which is how the census office divides Norway. These vary in population from zero to 3'500, with a mean of 320. The geocoding of the 64'240 daily trips and 27'507 long distance journeys involved two automatic matches and two manual correction phases against a set of address databases, including one with the names of firms and organisations (Denstadli and Hjorthol, 2003).

2.2 Swiss national travel survey (Mikrozensus 2000)

The Mikrozensus 2000 was the sixth in a series dating back to 1974 and is conducted by the Swiss Federal Office of Statistics (BFS) and the Federal Office of Spatial Planning (ARE) (2001 and 2002). A number of cantons provided additional support by financing additional respondents at marginal costs. The CATI-interview covered the stages of one entire day, and long distance and air travel for longer periods. The feasibility of geocoding the stage data was still uncertain during the survey's design phase, so exact street addresses or their equivalents were obtained only for trips to, within, and from the ten largest cities in Switzerland (40'000 to 340'000 inhabitants). The names of stations and public transport stops were carefully recorded as part of the stage-based interview, as well as home addresses. However, quality of address information was not a prime concern for the survey.

The geocoding (Jermann, 2003) of the 144'000 stages (about 100'000 trips[1]) was performed some time after the field phase of the survey, as part of a different project. Using geocoded address data bases of the BFS, canton Zürich, and the Swiss Federal Railways stations and stops, a semi-automatic matching process was implemented after normalising and correcting street addresses in the Mikrozensus 2000 records (spelling, punctuation, removal of diacritical marks etc.). The remaining addresses were matched by hand, as far as possible, using maps, telephone books, and information on the internet, especially for place names and leisure facilities. The address matching tools of the ArcInfo and MapInfo were unsuitable, as they embed too many assumptions valid only in an US context.

2.3 2003 Thurgau six-week diary

This survey replicates and improves on the 6-week Moboidrive survey (Axhausen, Zimmermann, Schönfelder, Rindsfüser and Haupt, 2002). A total of 99 households with 230 members were recruited in the rural and small town canton Thurgau; they reported their travel for a continuous six-week period, using six one-week trip diaries (about 36'000 trips). The data was coded on return and the field worker called back to clarify any omissions, particularly omitted or unclear addresses Address information quality was a clear priority for everyone involved in the survey.

The geocoding was undertaken (Machguth and Löchl, 2004) some time after the end of the field work using the same type of databases employed for the geocoding of the Mikrozensus 2000, and adopting the same process. In contrast to the Mikrozensus, destinations abroad were coded to street block level in Germany and to municipality level elsewhere.

3 Quality of geocoded locations

In the preceding section, we asked what level of quality could be achieved for such large-scale exercises when they rely primarily on automatic matching steps. The quality of geocodes can be evaluated by how precisely addresses can be pinpointed. In the Norwegian study, quality was rated by quantifying the number of wards to which an address could belong. Table 1 gives details on criteria for quality rankings. In nearly 90% of the cases, it

[1] Microzensus deliberately omitted many stages, in particular those under 100m; these omissions were exacerbated by interviewer error.

was possible to locate the address within one ward. However, address locations for both ends of the trip were possible in only 80% of the cases, raising problems later with distance calculations (Table 2). Trip purpose, mode, and area were investigated for impacts on accuracy. The first two were not significant, but the type of area, predictably, had an impact. Better databases for larger urban areas substantially improved quality, particularly when one considers that wards are smaller in these areas.

Table 1 2001 NPTS: Geo-information and accuracy level

Type of information	Accuracy level
1. Pre-geocoding of home address (verified by the respondent) 2. Pre-geocoding of work place address (verified by the respondent) 3. Street address, postal number, and municipality; location using GIS and address databases 4. As 3, but with some inaccuracies – manually controlled and verified 5. As 3, but using a manual method for location 6. Insufficient information (e.g., name of store, postal code etc.) but GIS or manual checks made possible exact location 7. Location to city centre in small urban settlements (few cases)	Exact location of statistical ward
7. As 6, but two possible wards	Approximate location (2 possible wards)
8. As 6, but three possible wards	Approximate location (3 possible wards)
9. As 6, but four or more possible wards	Inexact location (4 or more possible wards)
10. Insufficient information – only possible to locate municipality	No location
11. Geocoding impossible or destination abroad	No location

Table 2 2001 NPTS: Accuracy of the geocoded trip origins and destinations by area and by location

Accuracy of geocoding	Exact location of ward	Approximate location (2 or 3 possible wards)	Inexact location (4 or more possible wards)	Municipality only
Metropolitan areas of cities with 100'-500'000 inhabitants	81	4	10	5
Cities/towns of 40'-100'000 inhabitants	82	5	8	5
Smaller towns/villages	78	5	11	6
Sparsely populated areas	74	4	16	6
Trip origin	89	2	6	3
Trip destination	89	2	6	3
Origin and destination	78	4	11	6

The quality of Mikrozensus 2000 needs to be examined individually at each stage, as these were the basic trip unit descriptions.. Varying quality of underlying databases produces differences. Because some addresses were available only with street names, and in most cases only as municipalities, collection of addresses differed for various areas during the survey. Table 3 details the quality rankings and Table 4 the qualities available at stage ends.

Matching is very good for stages with stations at either end, relatively good for both other public transport stops, and categories. It is interesting how well street addresses could be coded, when they were available. However, in one third of the cases, respondents could only recall the street, or only a street could be identified for the location. The municipalities were matched precisely. Note that category C2, which refers to locations for available street addresses, was so incomplete that only a matching could only be achieved at the municipal level Slightly more than 70% of the stages could be matched at both ends to level 1 – (including 14% municipality to municipality stages) - and 85% to level 1 or 2, which is roughly comparable to the Norwegian results. Considering that the average Swiss municipality has only about 2500 inhabitants, and given that the Mikrozensus was mostly conducted without considering geocoding of locations, this is a very good result.

Table 3 Mikrozensus 2000: Rating of the matching quality by type of location

Rating	Description	Quality
Building address available		
A1	Precise match	Precise
A2	Varying address spelling, certain match	Certain
A3	Strongly varying spelling, uncertain match	Uncertain
Street name available		
B2	No house number available; employed lowest known number in the street	Certain
B3	As above, but uncertain match	Uncertain
Municipality known		
C1	No street address	Precise
C2	Street address given, but not identifiable locally	Certain
C3	Dubious information in the Mikrozensus	Uncertain
Bus or tram stop		
D1	Precise match	Precise
D2	Varying address spelling, certain match	Certain
D3	Strongly varying spelling, uncertain match	Uncertain
Station		
E1	Precise match	Precise
E3	Strongly varying spelling, uncertain match	Uncertain
F	Not identifiable; abroad	No match

Table 4 Mikrozensus 2000: Matching quality by stage end

	From														
To	A1	A2	A3	B2	B3	C1	C2	C3	D1	D2	D3	E1	E3	F	Sum
A1	4.0	0.4	0.0	2.6	0.1	3.4	0.1	0.0	1.5	0.3	0.1	6.0	0.0	0.7	19.3
A2	0.4	0.2	0.0	0.2	0.0	0.8	0.0	0.0	0.1	0.0	0.0	1.0	0.0	0.1	3.0
A3	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	-	0.0	-	0.0	0.1
B2	2.4	0.2	0.0	1.9	0.0	1.0	0.0	0.0	0.5	0.1	0.0	1.3	0.0	0.2	7.9
B3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	-	0.0	0.0	0.0	0.0	-	0.0	0.2
C1	3.2	0.7	0.0	0.9	0.0	12.2	0.3	0.0	0.2	0.1	0.0	4.0	0.0	0.4	22.1
C2	0.1	0.0	0.0	0.1	0.0	0.3	0.1	0.0	0.0	0.0	0.0	0.4	0.0	0.0	1.0
C3	0.0	0.0	-	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.1
D1	1.5	0.1	0.0	0.5	0.0	0.2	0.0	0.0	1.7	0.2	0.1	0.8	0.0	0.1	5.3
D2	0.3	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.2	0.1	0.0	0.2	0.0	0.0	1.1
D3	0.1	0.0	-	0.1	0.0	0.0	0.0	-	0.1	0.0	0.0	0.1	0.0	0.0	0.5
E1	5.7	1.0	0.0	1.2	0.0	4.1	0.4	0.0	0.9	0.2	0.1	21.4	0.1	0.4	35.5
E3	0.0	0.0	0.0	0.0	-	0.0	0.0	-	0.0	0.0	0.0	0.1	0.1	0.0	0.3
F	0.7	0.0	0.0	0.2	0.0	0.4	0.0	0.0	0.1	0.0	0.0	0.6	0.0	1.5	3.7
Sum	18.5	2.8	0.1	7.9	0.2	22.6	0.9	0.1	5.3	1.1	0.5	36.0	0.3	3.6	100.0

The geocoding quality for 2003 Thurgau followed the Mikrozensus example, but was supplemented by a new coding that translated the previous codes into more comprehensible metric (see Table 5). The code “up to 100m” is understating the accuracy, as it concerns mainly exactly coded street addresses. The quality of the geocoding is very high, reflecting the attention given to it during the survey process. With 60% of trips captured within 100 m of their true origins and destinations, one is very close to ideal conditions for the later distance estimation.

Table 5 2003 Thurgau: Matching quality by trip end

Quality at origin	Quality at destination					
	100 m	500 m	1000 m	Municipality	Unknown	Sum
100 m	60.3	13.4	0.1	2.7	0.6	77.1
500 m	13.4	3.2	0.0	0.9	0.1	17.6
1000 m	0.1	0.0	0.0	0.0	-	0.3
Municipality	2.6	1.0	0.0	0.6	.0.0	3.2
Unknown	0.6	0.1	-	0.0	0.0	0.7
Sum	77.0	17.7	0.2	4.2	0.7	100.0

4 Differences between distance estimates

Swiss and Norwegian data allow comparison of network estimates against reported distances, as well as against each other. This section focuses on the comparison between the various network estimates discussed above.

In a first step for Mikrozensus 2000, the stage based information discussed above was used to geocode the trips. The best available geocode was attached to the start of the first stage and the destination of the last stage (See Table 6). The main mode of the trip was determined, as usual in this situation, by an a-priori ranking of the modes involved, in which the various public transport modes have priority before private motorised vehicles and slow modes. Further analysis in this section was restricted to car driver and passenger trips, as no detailed walking and cycling network information was available.

Network distance calculations were performed using a national assignment model available at IVT (Vritc, Fröhlich and Axhausen, 2003 or Vritc and Axhausen, 2004), which breaks Switzerland down into 3’066 zones, 14’798 nodes, and 19’664 links. The associated origin-destination matrix of average annual weekday flows is calibrated for the year 2001. The geocode for a post code is the geocode of the associated post office’s address.. As a municipality is normally the same as a post code area and a zone in the national network model, this address was also used to describe the centre of gravity of the zones. Distance between the network and centre of gravity, i.e. the length of centroid connector, was set to the Euclidian distance between the relevant node and the centroid.

Table 6 Mikrozensus 2000: Quality of the geo-coding of trip origins and destinations (104'215 trips; all modes)

Trip origin	Trip destination			
	Post code, street name and house number	Post code and Street name	Only post code	Total
Post code, street name and house number	16.8	0.0	6.2	23.0
Post code and Street name	0.0	0.0	6.3	6.3
Only post code	0.0	0.0	70.7	70.7
Total	16.8	0.0	83.2	100.0

Crow-fly distances are calculated as Euclidian distances between the origin and destination of the trip, at the precision available. For network-based calculations, each trip end was associated with the relevant zone, and therefore its centroid. Distances were calculated using VISUM 8.0 (PTV AG, 2002). Shortest distance path distances include lengths of centroid connectors at either end of trips. Shortest-time path distances were calculated assuming free-flow speeds for links. User-equilibrium (UE) assignment distances were calculated as weighted average distances of paths used at equilibrium between any two locations. The matrix of average weekday traffic flows was assigned with the assumption that daily link capacities are twelve times hourly link capacities. All trips inside a zone were excluded from further analysis, as they have, by definition, a distance of zero in network models, better interpreted as a missing value.

Comparison of distance distributions (Table 7 and Figure 1) highlights differences between the three sources of information. Crow-fly distances have their mode in the 1-5 km band and smaller shares for all successive bands. Mean crow-fly distance is substantially smaller then other means. Network distance distributions are similar, but, as one would expect, shortest-time path and mean UE assignment path distances are slightly longer. This effect is pronounced for longer distances, where routings via roads with higher speeds start to pay off. Alpine topography, including the many large lakes in the foothills of the Alps, explains the large differences in shares of trips over 100 km distance vs. crow-fly distances. Mean reported distance lies between the shortest-distance path and shortest-time path estimate. Given that neither of the two network-based estimates reflects actual behaviour fully, this mean value is a credible estimate for all trips. Wolf et al., 2003 support this conclusion by showing that their GPS traced distances are, depending on time of day, about 10% shorter than UE distance estimates.

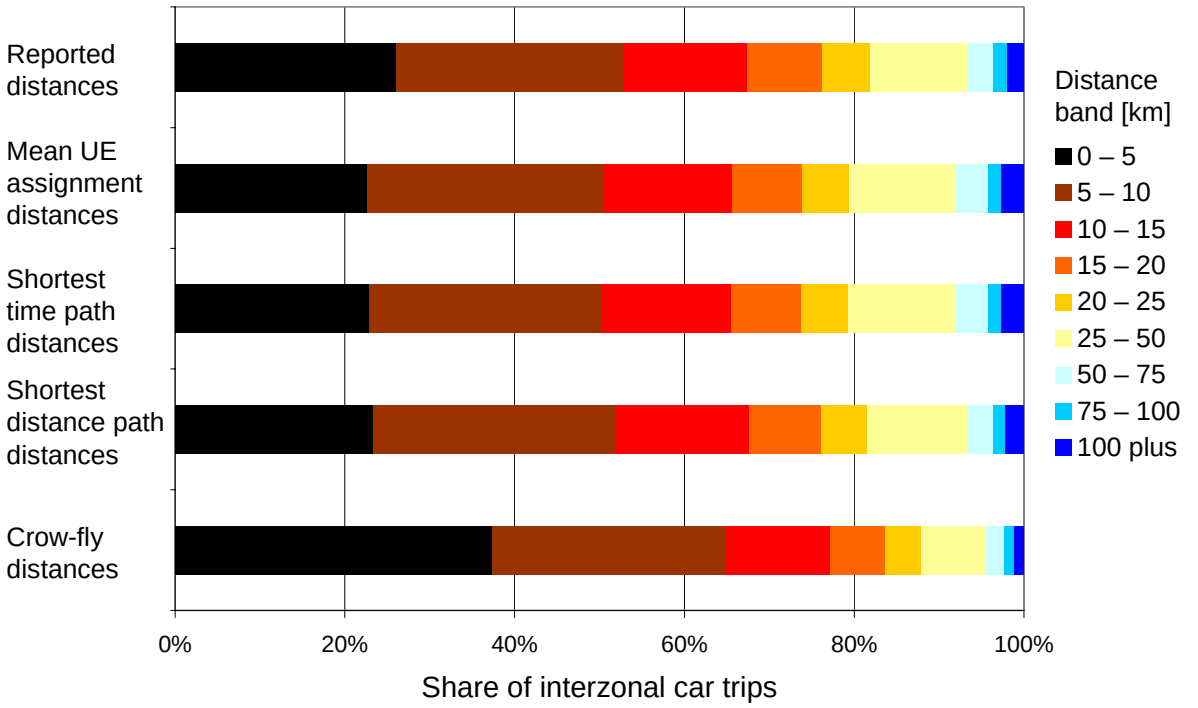
Table 7 Mikrozensus 2000: Distribution of the reported and calculated distances (34'195 car passenger and driver interzonal trips)

Distance band [km]	Crow-fly		Shortest distances path		Shortest time path		Mean UE paths		Reported	
	Share [%]	Class mean [km]	Share [%]	Class mean [km]	Share [%]	Class mean [km]	Share [%]	Class mean [km]	Share [%]	Class mean [km]
0 – 5	37.34	3.1	23.32	3.5	22.93	3.5	22.59	3.5	26.10	3.6
5 – 10	27.45	7.1	28.60	7.3	27.35	7.3	27.76	7.3	26.65	7.9
10 – 15	12.32	12.3	15.77	12.3	15.18	12.3	15.19	12.4	14.67	13.1
15 – 20	6.55	17.3	8.38	17.3	8.34	17.2	8.36	17.2	8.77	18.3
20 – 25	4.26	22.4	5.51	22.3	5.54	22.2	5.51	22.2	5.72	23.3
25 – 50	7.63	34.2	11.69	34.1	12.75	34.2	12.66	34.2	11.40	35.5
50 – 75	2.20	60.4	3.17	61.0	3.66	61.1	3.64	61.0	3.05	62.0
75 – 100	1.07	85.7	1.37	86.22	1.60	86.9	1.63	86.6	1.71	88.6
100 plus	1.18	135.2	2.20	148.2	2.67	161.0	2.65	161.5	1.94	158.7
Total	100.0	13.1	100.0	17.9	100.0	19.6	100.0	19.6	100.0	18.4

In many cases, it is useful to convert one distance estimate to another. Such conversion or detour factors have been previously reported, but only for certain pairs of distance estimates (for example by Qureshi, Hwang, and Chin, 2002). Table 8 provides six comparisons for Mikrozensus 2000 based on the estimates described above. A clear difference can be observed in detour factors change patterns. . Calculations are based on all observations in the sample, even if crow-fly distances were longer than model based estimates. This can happen, especially for shorter trips, when the distance between zonal centroids is smaller than actual distance travelled (see above). Detour factors fall as crow fly distances become longer. While they are well above the square root of two – factor of the Manhattan metric for short distances, they are also much smaller for longer distances. Factors for the three network distances are, for practical purposes, identical for the shortest distance band, but diverge after this, reflecting different objective functions behind their calculation.

The pattern is reversed for shortest distance paths detour factors, where the factors grow with increasing shortest path distance. This is predictable, as the chance to use a faster, but longer route via the sparser high capacity network increases with trip length.

Figure 1 Mikrozensus 2000: Comparison of the distance distributions (34'195 car passenger and driver interzonal trips)



In the 2003 Thurgau survey, the distances (shortest distance and shortest time path) were calculated using high resolution Vektor 25 – network of the Swiss ordinance survey, employing the gecodes described above. This allowed the inclusion of all trips, but for cases where respondents return to the same address after a walk or drive. The patterns revealed in Table 9 are similar to those discussed for the Mikrozensus 2002, but their levels are markedly lower for crow-flow distance ratios , reflecting the finer network employed and the absence of centroid connectors.

Distance estimate comparisons for the Norwegian data are possible only for shortest time path distance at this time. However, results confirm the pattern revealed by the Mikrozensus data; the detour factor is significantly larger in the shortest distance band (Table 10). The national level planning network was provided by the Norwegian highway authority and the path calculation included travel times, distances, and various bridge and ferry tolls.

Table 8 Mikrozensus 2000: Detour factors between different distance estimates (34'195 car passenger and driver interzonal trips)

Average detour factor with	Crow fly distance			Shortest distance paths		Shortest time distance
Distance band	Shortest distance path distance	Shortest time path distance	Mean user equilibrium distance	Shortest time path distance	Mean user equilibrium distance	Mean user equilibrium distance
0 to 5 km	1.83	1.87	1.88	1.01	1.02	1.01
5 to 10 km	1.39	1.46	1.46	1.04	1.05	1.00
10 to 25 km	1.35	1.47	1.47	1.09	1.09	1.00
25 to 50 km	1.31	1.46	1.46	1.11	1.11	1.00
50 to 75 km	1.31	1.47	1.47	1.12	1.12	1.00
75 to 100 km	1.32	1.49	1.49	1.13	1.13	1.00
100km and more	1.26	1.48	1.48	1.16	1.16	1.00
Total	1.54	1.62	1.62	1.05	1.05	1.00

Table 9 2003 Thurgau: Detour factors between different distance estimates (car passenger and driver; public transport; slow modes)

Average detour factor with	Public transport			Car driver and passenger			Slow modes		
	Crow fly distances		SDPD	Crow fly distances		SDPD	Crow fly distances		SDPD
Distance band	SDPD	STPD	SDPD	STPD	SDPD	STPD	SDPD	STPD	STPD
0 to 5 km	1.33	1.38	1.05	1.46	1.50	1.04	1.44	1.49	1.04
5 to 10 km	1.46	1.51	1.02	1.35	1.40	1.02	1.67	1.73	1.01
10 to 25 km	1.26	1.32	1.05	1.25	1.32	1.05	1.81	1.85	1.03
25 to 50 km	1.20	1.32	1.10	1.21	1.32	1.09			
50 to 75 km	1.25	1.40	1.09	1.26	1.39	1.08	1.26	1.36	1.08
75 to 100 km	1.30	1.43	1.12	1.30	1.46	1.12			
100 km plus	1.28	1.34	1.07	1.19	1.29	1.11			
Total	1.28	1.36	1.06	1.38	1.43	1.04	1.45	1.50	1.04

SPDP: Shortest distance path distance; STPD: Shortest time path distance; all values shown are based on 30 or more observations.

Table 10 2001 NPTS: Mean and median detour factors between shortest time path distance and crow fly distance by distance band (20'700 car passenger and driver trips below 100 km)

	Detour factor	
Distance band	Mean	Median
0-9 km	1.56	1.48
10-19 km	1.42	1.34
20-29 km	1.40	1.33
30-39 km	1.37	1.32
40-49 km	1.40	1.36
50-> km	1.43	1.35
Total	1.51	1.42

Figure 3 Ratios of shortest time paths with crow fly distances by distance band

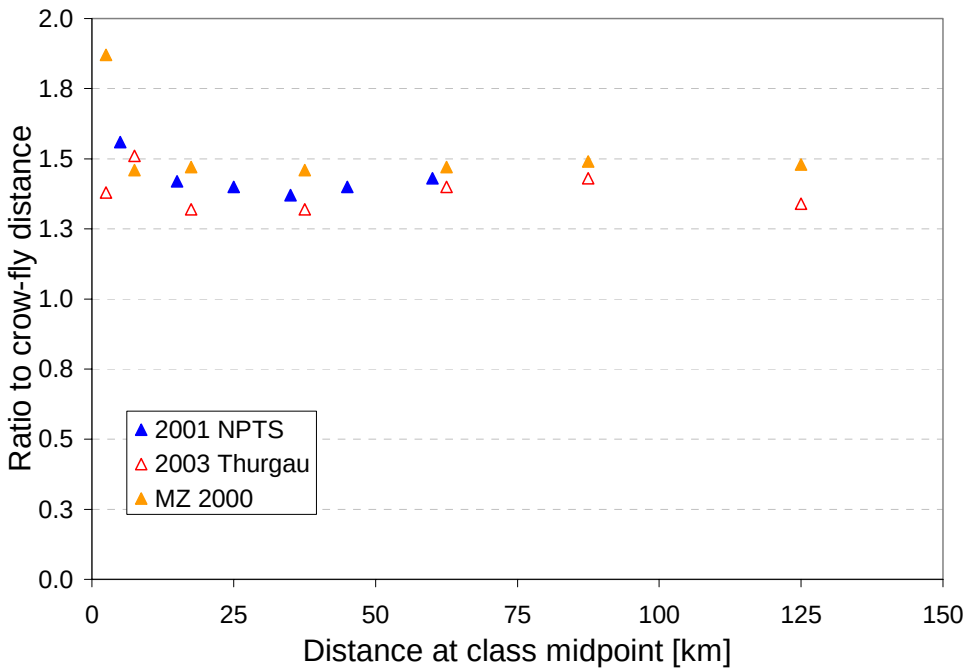


Figure 3 visualises and compares results for shortest-time path distances. Ratio level seems to depend on resolution of the networks used. The national level planning networks used for

Mikrozensus 2000 and 2001 NPTS produce bigger ratios than the finer network used for the 2003 Thurgau survey. This is especially obvious for the shorter distance bands, while differences start to disappear over long distances. .

5 Reported and estimated distances

Unknown errors associated with travellers' reported distance estimates have led modellers to avoid their use wherever possible. Expressly, when estimating choice models, the consistent errors of network models are preferable to travellers' unknown, idiosyncratic errors. But, in many cases, neither geocodes nor network models are available. Thus, the quality of reported distances is important, especially if the errors were to cancel out for averages or other sample summaries.

One way to assess reported distance quality is to compare it to the shortest distance path distance derived from a network model. If zone based, one can asses the measurement uncertainties associated with inter-zonal distances compared with distances between addresses. In the 2001 NPTS, geocodes refer to statistical wards of differing size. To ascertain measurement uncertainty, mean distance between every ward address and its centroid was calculated for each ward (for details see Denstadli and Engebretsen 2004). To avoid large measuring uncertainties, trips to/from wards with mean distance more than 1.0 km were eliminated. In addition, trips with obvious geocoding errors and and trips where the measurement uncertainty for either statistical ward was larger than one quarter of the network distance estimate were removed. Finally, trips that started and ended in the same ward were omitted.

The resulting relative errors are shown in Table 11 by distance band for all car driver and passenger trips below 100 km, the vast majority of all such trips. The measurement uncertainty is nearly independent of trip distance and fairly small, with a mean of about 0.6 km.

Table 11 2001 NPTS: Distribution of the relative errors of reported to shortest time path distance estimate by distance (20'700 car passenger and driver trips below 100 km)

Shortest time path distance	Share of trips with relatives error of reported to shortest distance path distance estimate [%]						
	Within the measuring uncertainty	< 5 %	< 10 %	< 25 %	< 50 %	50 % +	Total
0-9 km	28,2	8,7	8,2	19,5	19,1	16,3	100,0
10-19 km	17,1	15,9	13,2	23,8	15,9	14,1	100,0
20-29 km	12,6	18,0	18,1	26,4	14,1	10,7	100,0
30-39 km	13,4	24,3	14,8	22,6	12,1	12,8	100,0
40-49 km	9,5	23,9	27,6	22,1	5,5	11,3	100,0
50-> km	7,1	25,5	18,4	24,1	9,4	15,5	100,0
Total	23,6	12,0	10,7	21,1	17,4	15,3	100,0

The overall error decreases with distance. The shares of trips in the various error bands are redistributed. The large share of distance estimates within the measuring uncertainty is noticeable for the lowest distance band. This share goes down with distance with a nearly matching increase in the below 5% error band. About 45% of trips are estimated within 10% of the shortest time path distance. Additional analysis showed minor difference between different trip purposes, young and middle-aged people, sexes, and between urban and rural areas.

Errors in reported distances are not due only to respondents misinforming, but may also be caused by interviewer misinterpretation or recording errors. We expect errors of this kind to be more random. Plots of reported distances against distances from the network model show that, except for some outliers, distance estimates are highly correlated. Omitting the outliers, we can conclude that deviations seem randomly and asymptotically normally distributed (for details see Denstadli and Engebretsen, 2004), with the result that the mean detour factor is close to 1.0 for all distance bands (Table 12).

Repeating this analysis for the 2000 Mikrozensus data (Table 13) also reveals a similar pattern for public transport trips. Mean detour factors are dominated by outliers over short distances. Over longer distances, the median converges quickly to one for car trips and to 1.1 for longer public transport trips. The factor drops below 1.0 for longer car trips and to about 1.2 for public transport trips. To obtain a credible estimate of distance travelled this pattern

requires adjustment of reported distances by distance band. The poorer estimates for public transport reflect the longer routing of public transport services, a lack of active navigation by the traveller, and slow access and egress to the station or stop.

Table 12 2001 NPTS: Detour factors between reported and shortest time path distance (20'700 car passenger and driver trips below 100 km)

Shortest time path distance	Detour factors	
	Mean	Median
0-9 km	1.11	0.96
10-19 km	0.99	0.99
20-29 km	1.00	1.03
30-39 km	0.96	1.02
40-49 km	0.99	1.02
50-> km	0.91	1.01
Total	1.07	0.99

The pattern is also visible in Thurgau 2003, but not as clearly. It is obvious that the very large detour factors for short distances in Mikrozensus 2000 data are a product of omitted intra-zonal trips. The very low reported distances in the longer distance band are due to the omission of hiking and cycling paths in the network model used; these can be crucial in hilly terrain. It should be noted that the speed assumptions chosen for shortest time paths were overly optimistic resulting in reported travel time underestimates of about 1/3. . This is far too much, even allowing for errors inherent in reported travel times. One would assume that this would lead to longer-than-realistic distances for longer trips.

The pattern of change suggests a relationship with trip speed and its mode. Based on the distance bands used above, this hypothesis is confirmed by Figure 4. The same pattern, but without the outlier for the short interzonal distances, can be seen in the 2003 Thurgau data.

For the Mikrozensus 2000 data, which represent amore typical situation, the dependence of the detour factor on the reported speed was modelled using aggregate values for distance bands of 2 km up to 50km and of 5 km beyond that. The best fitting model is shown in Table 15. For an alternative approach, see Zmud and Wolf, 2003.

Table 13 2000 Mikrozensus: Detour factors between reported and shortest distance path distance (car passenger and driver and public transport interzonal trips)

Average detour factor with	Public transport		Car driver and passenger	
Distance band	Mean	Median	Mean	Median
0 to 5 km	4.123	3.339	1.584	1.204
5 to 10 km	1.590	1.554	1.156	1.019
10 to 25 km	1.437	1.282	1.074	1.000
25 to 50 km	1.177	1.036	1.049	0.997
50 to 75 km	1.167	1.073	0.991	1.008
75 to 100 km	1.106	1.145	0.940	0.999
100km and more	1.164	1.176	0.825	0.985
Total	1.484	1.225	1.205	1.032

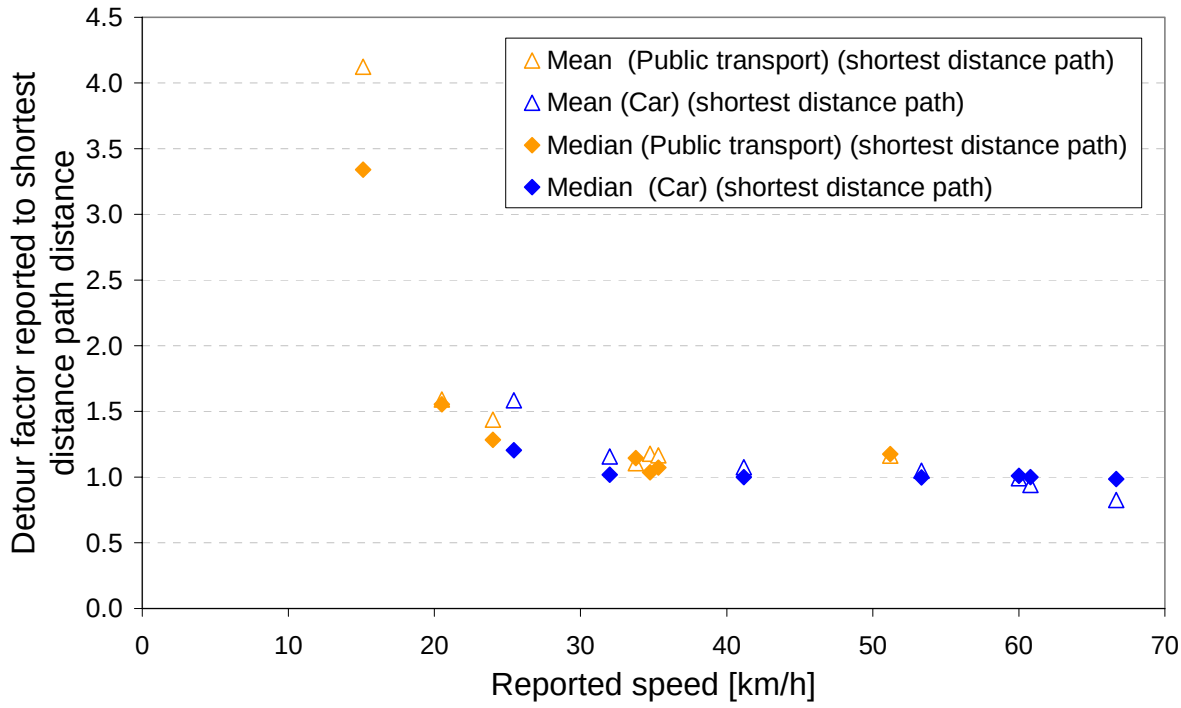
Table 14 2003 Thurgau: Detour factors between reported and shortest distance path distance

Average detour factor with	Public transport		Car driver and passenger		Slow modes	
Distance band	Mean	Median	Mean	Median	Mean	Median
0 to 2.5 km	1.32	1.16	1.16	1.07	1.17	1.04
2.5 to 5 km	0.97	1.01	1.03	1.02	0.81	0.92
5 to 10 km	1.20	1.20	1.12	1.07	0.90	1.11
10 to 25 km	1.15	1.15	1.10	1.13	0.65	0.10
25 to 50 km	1.01	1.11	1.02	1.09		
50 to 75 km	1.10	1.17	1.14	1.13	0.33	0.06
75 to 100 km	1.12	1.16	1.04	1.08		
100km and more	1.13	1.14	1.10	1.06		
Total	1.32	1.16	1.16	1.07	1.17	1.04

Table 15 Mikrozensus 2000: Linear regression of detour factors between reported and shortest distance path distance on reported speed

Variable	Parameter	t-Value
Constant	-1.940	-3.129
Reported speed	.771	6.234
Inverse reported speed * 100	.028	3.707
N	118	
Adjusted R2	0.491	

Figure 4 Distributions of reported distance deviations from calculated distances



6 Conclusions and further research

The three questions raised at the beginning of this paper were:

- How accurate can the geocoding of addresses obtained from travel diaries be?
- How big are the differences between various distance estimates?
- What are the differences between reported distances and calculated distances?

The experiences reported here show that, in urban areas, it is possible to geocode almost all locations to within 100 m of their true geocode, if the survey process emphasises this aspect of the work. With lower accuracy requirements, higher rates are possible. This carries forward in the joint accuracy of the trip length estimate, as the probability increases that both trip ends are well coded. It should be noted, though, that these rates require very good address databases, especially for firms, commercial outlets, common locations without street addresses, and public transport stations and stops. The last two categories require particular attention, as these addresses are often not available from either the relevant Census office or commercial providers. (In the case of Norway and Switzerland, it was possible to obtain relevant databases from public transport operators or the national government) National public transport timetables do include some geocoding information, but their station and stop names sometimes differ from local nomenclature.

Lower location rate for trips undertaken outside urban areas (noticeable in the 2001 NPTS, as well as other surveys), raises some concern. The low location rate is due to a lack of street names and identifiable landmarks like shops, churches, etc. It is important that the interviewer keeps this in mind. If the respondent is unable to provide an address or a landmark close by, the interviewer must make him/her describe the place in alternative ways, e.g. by asking for distance and direction to the nearest lake or urban settlement, or any other marker that can help locate the trip.

There are large and systematic differences in network distance estimates, as expected. It is crucial that the modeller reports the assumptions behind the estimates used. The 2003 Thurgau data shows that speed assumptions behind the shortest-time path distances can be crucial; detour factors provided here give a first impression of their size and pattern. However, they cannot be corroborated until the literature provides further estimates of their value. Still, the impact of network resolution is already visible in the results reported here.

Differences between reported and estimated distances can be very large for an individual trip. These errors do not cancel out for large samples. A systematic difference remains, but its pattern is predictable and depends on the trip distance. For longer trips, the medians of reported distances match the shortest-distance path distances. Correcting for reported speed, there are no differences in detour factors between modes. The strong dependence on reported speed suggests a reasonable way to correct estimates.

Although we do not recommend using self-reported information as the only data for travel distances, self-reported distances are useful when assessing the quality of geocoding. Large deviations between two distance measures may indicate that it is an incorrectly located start or

end point and not the respondent's stated travel distance. There may also be errors in digital road data or logical defects in models determining the route (and consequently the distance). In addition, as long as objective measurements relate only to distances between zones (e.g. statistical wards), self-reported distances represent valuable additional information on short trips and intra-zone trips.

Three surveys do not allow wide generalisations. Replication of this work is required to establish the robustness of the results presented here. Discrepancies due to different formulations of networks models are especially important, as there is substantial variance in professional practise, which should be reduced to improve accuracy and consistency of the model results. This zeros in on the most important element missing for further research: a high-quality GPS dataset matched to an equally high quality network model as the basis for detailed studies.

7 Acknowledgements

The authors are grateful for the support of M. Machgut and J. Jermann during the geocoding of the Swiss data and for the support of Mr. M. Vrtic and Mr. T. Hamre, who provided the network distance estimates for the 2000 Mikrozenus and the 2001 NPTS respectively.

The results are our own and do not reflect the assessment of the owners of the datasets used.

8 References

Axhausen, K.W. (2003) Definitions and measurement problems, in K.W. Axhausen, J.L. Madre, J.W. Polak and P. Toint (eds.) *Capturing Long Distance Travel*, 8-25, Research Science Press, Baldock.

Axhausen, K.W., A. Zimmermann, S. Schönfelder, G. Rindsfüser and T. Haupt (2002) Observing the rhythms of daily life: A six-week travel diary, *Transportation*, **29** (2) 95-124.

Axhausen, K.W., J.L. Madre, J.W. Polak and P. Toint (eds.) *Capturing Long Distance Travel*, Research Science Press, Baldock.

Bovy, P.H.L. and E. Stern (1990) *Route Choice: Wayfinding in Transport Networks*, Kluwer, Dordrecht.

Bundesamt für Raumentwicklung, Bundesamt für Statistik (2001) Mobilität in der Schweiz, Ergebnisse des Mikrozensus 2000 zum Verkehrsverhalten, Bern und Neuenburg.

Bundesamt für Raumentwicklung, Bundesamt für Statistik (2002) Mikrozensus Verkehrsverhalten 2000, Hintergrundbericht zu „Mobilität in der Schweiz“, Bern und Neuenburg.

Denstadli, J.M. and R.J. Hjorthol (2003) Testing the accuracy of collected geoinformation in the Norwegian Personal Travel Survey – experiences from a pilot study, *Journal of Transport Geography*, **11** (1) 47-54.

Denstadli, J.M., R. Hjorthol, A. Rideng and J.I. Lian (2003) Travel behaviour in Norway, TØI report, 637/2003, Institute of Transport Economics, Oslo.

Denstadli, J.M. and Ø. Engebretsen (2004) Testing the accuracy of self-reported geoinformation travel surveys, paper submitted to the Conference on progress in activity-based analysis, Maastricht, 28-31 May 2004.

Hackney, J., F. Marchal and K.W. (2004) Monitoring a road system’s level of service: The Canton Zürich floating car study 2003, paper submitted for presentation at the 84th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2005.

Hubert, J.P. (2003) GIS-based enrichment, in K.W. Axhausen, J.L. Madre, J.W. Polak and P. Toint (eds.) *Capturing Long Distance Travel*, 256-278, Research Science Press, Baldock.

Jermann J. (2003) Geokodierung Mikrozensus 2000, *Arbeitsbericht Verkehrs- und Raumplanung*, **177**, IVT, ETH Zürich, Zürich.

Machguth, H. und M. Löchl (2003) Geokodierung 6-Wochenbefragung Thurgau 2003, *Arbeitsbericht Verkehrs- und Raumplanung*, **219**, IVT, ETH Zürich, Zürich.

Marchal, F., J.K. Hackney and K.W. Axhausen (2004) Efficient map-matching of large GPS data sets - Tests on a speed monitoring experiment in Zurich, paper submitted to the 84th Annual Meeting of the Transportation Research Board, Washington, Janaury 2005.

Ortuzar, J. de D. and L.G. Willumsen (2001) *Modelling Transport*, Wiley, Chichester

PTV AG (2002) User Manual VISUM 8.0, Planung Transport Verkehr AG, Karlsruhe.

Qureshi M. A., H. Hwang, and S. Chin (2002) Comparison of distance estimates for the commodity flow survey based on the great circle distance versus network based distances, *Transportation Research Record*, **1804**, 212-216.

Raghubir, P., and A. Krishna (1996) As the crow flies: Bias in consumers’ map-based distance judgments, *Journal of Consumer Research*, **23** (1) 26-39.

Resource Systems Group (1999) Computer-based intelligent travel survey system: CASI/Internet travel diaries with interactive geo-coding, report to the U. S. Department of Transportation, RSG, Manchester.

Richardson, A.J., E.S. Ampt and A.H. Meyburg (1995) *Survey Methods for Transport Planning*, Eucalyptus Press, Melbourne.

Rietveld P., B. Zwart, B. Van Wee and T. van den Hoorn (1999) On the relationship between travel time and travel distance of commuters: Reported versus network travel data in the Netherlands, *The Annals of Regional Science*, **33** (3) 269-287

Sheffi, Y. (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Englewood Cliffs.

Vrtic, M. and K.W. Axhausen (2004) Forecast based on different data types: A before and after Study, paper presented at 10th World Conference on Transport Research, Istanbul, July 2004.

Vrtic, M., P. Fröhlich and K.W. Axhausen (2003) Schweizerische Netzmodelle für Strassen- und Schienenverkehr, in T. Bieger, C. Laesser and R. Maggi (eds.) *Jahrbuch 2002/2003 Schweizerische Verkehrswirtschaft*, 119-140, SVWG, St. Gallen.

Wolf, J., M. Oliveira and M. Thompson (2003) The impact of trip underreporting on VMT and travel time estimates: Preliminary findings from the California statewide household travel survey GPS study, paper presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., January 2003.

Zmud, J. and J. Wolf (2003) Identifying the correlates of trip misreporting: Results from the California statewide household travel survey GPS study, paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003.